

--	--	--	--	--	--	--	--	--	--	--	--

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2016/2017

TDS2101 – INTRODUCTION TO DATA SCIENCE
(All sections / Groups)

25 FEBRUARY 2017
9.00 a.m. - 11.00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 pages with 4 Questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

Question 1

a) State THREE key-enablers for the growth of big data. (1.5 marks)

b) You have been tasked to explain the characteristics of Big Data to a newcomer to the data science team. Explain briefly the 5V's that are commonly used to characterize Big Data. (2.5 marks)

c) There are several challenges faced by Data Scientists within a Data Science project life cycle. What are FOUR possible challenges that may affect the project? (2 marks)

d) Describe FOUR structures of data commonly found in Big Data project datasets and give ONE example for each. (4 marks)

Question 2

a) The Fit n Sleek company is about to launch a new product e.g. Sleep Tracker. Ben, who is a Data Scientist, was tasked to determine potential customers who might purchase the product. Ben consulted a sleep expert who suggested that an increase in sleep hours might reduce daytime drowsiness.

- Among the six *types of questions* that can be asked to direct the analysis, determine which *type of question* would be most appropriate and state why.
- Formulate the question based on the type determined in (i)

(2 marks)

b) To qualify for a PhD scholarship offered by an Australian graduate school, a candidate must not be more than 45 years old on 1st January 2017. In an effort to filter qualified candidates for their PhD scholarship, the graduate school decided to examine the date-of-birth (D.O.B) column in their database. Upon conducting a random manual check of the application form and the candidates ID card, it was discovered that some candidates had entered their date-of-birth in the US date format i.e. MM/DD/YYYY instead of the expected DD/MM/YYYY format.

- What impact would this have on the graduate school's scholarship offer?
- Which data quality dimension does this violate?
- How can this problem be avoided in the future?

(3marks)

Continue...

c) The dataset below shows a subset of relative consumption of certain food items in European and Scandinavian countries. The values represent the percentage (%) of the population consuming that food type.

	Country	1	2	3	4	5
		Real coffee	Instant coffee	Tea	Sweetener	Biscuits
1	Germany	90	49	88	19	57
2	Italy	82	10	60	2	55
3	France	88	42	63	4	76
4	Holland	96	62	98	32	62
5	Belgium	94	38	48	11	74
6	Luxembourg	97	61	86	28	79
7	England	27	86	99	22	91
8	Portugal	72	26	77	2	22
9	Austria	55	31	61	15	29
10	Switzerland	73	72	85	25	31
11	Sweden	97	13	93	31	
12	Denmark	96	17	92	35	66
13	Norway	92	17	183	13	62
14	Finland	98	12	84	20	64
15	England	27	86	99	22	91
16	Spain	70	40	40		62
17	Ireland	30	52	99	11	80

- Data cleaning is an integral step in the data science pipeline. Identify THREE errors/inconsistencies that occur in the dataset above. State explicitly where the problem occurs (e.g. indicate the row and column number).
- Among the errors/inconsistencies identified in (i), suggest and justify an approach to handle ONE of the errors/inconsistencies identified.

(5 marks)

Question 3

a) The Players raw data given below shows the first name, gender, weight and height of state athletes.

```
raw_data = {'first_name': ['Harry', 'Lisa', 'Amy', 'Edward', 'Elise'],
'gender': [1, 0, 0, 1, 0],
'weight': [65, 60, 70, 75, 55],
'height': [172, 156, 160, 171, 159]}
```

- Write a Python code to load the data into a Pandas data frame and then compute the correlation coefficient value between the two continuous

variables of the Players dataset. Note: Be sure to include the import statements as necessary.

ii. What does a negative correlation coefficient value imply?

(4 marks)

b) Will, the Director of MeFit Sdn. Bhd, has collected a large amount of data from forum conversations on his company website. He would like to know what are the different topics being discussed by his customers. Would a supervised or unsupervised machine learning approach be more suitable for this purpose? Justify your choice.

(2 marks)

c) Edwin, a Data Scientist, collected data pairs (age, height) of humans from birth to the age of 60 years. If a scatterplot is drawn with this dataset,

- should he anticipate a positive or negative correlation? Justify your answer.
- would it show a linear or non-linear relationship? Justify your answer.

(4 marks)

Question 4

a) Describe THREE architectural structures of data storage in NoSQL databases and give one example of a NoSQL database

(4 marks)

b) Suggest TWO reasons why Python may be a preferred programming language for Data Science projects.

(4 marks)

c) The Chief Data Scientist at WordGadgets has conducted extensive analysis to determine ways to enhance their marketing campaign to boost sales. Suggest an appropriate chart/graph that can be presented to convey the following:

- Relationship between sales volume and product price
- Distribution of product sales across the states

(2 marks)

End of Paper...